ニュース事象の5Wモデルとそれに基づく事象関係分析

AUTHOR(S):

堀江, 伸太朗; 切通, 恵介; 馬, 強

CITATION:

堀江, 伸太朗 ...[et al]. ニュース事象の5Wモデルとそれに基づく事象関係分析. DEIM2016 2016: C3-4.

ISSUE DATE:

2016-03

URL:

http://hdl.handle.net/2433/217594

DEIM Forum 2016 C3-4

ニュース事象の5Wモデルとそれに基づく事象関係分析

堀江伸太朗† 切通 恵介† 馬 強†

† 京都大学大学院情報学研究科 〒6068501 京都府京都市左京区吉田本町
E-mail: †{horie,kiritoshi}@db.kyoto-u.ac.jp, ††ma@kyoto-u.ac.jp

あらまし ニュース記事には，事象の由来や背景について簡潔にまとめて記述されることがあり，読者にとって理解が困難である場合がある．これらの簡略化された抽象事象に対応する，より詳細的な具体事象の提示がユーザの理解支援には有効である．本研究では，ニュース記事中で言及されている事象を Who，Whom，What，When，Where の 5W 要素で表現し，それに基づく事象の分析手法を提案する．提案手法では，まず，ニュース記事から 5W 要素を抽出する難易度及びそれぞれの要素の抽象度に基づいて事象の抽象度を測る．そして，事象間の類似度を計算し，抽象度によって類似する事象間の抽象-詳細の対応関係を分析する．

キーワード ニュース，抽象度分析，5W モデル，事象間関係分析

1. はじめに

近年、Web 上で配信されるニュース記事は膨大な数に上っており、その中から有用な記事を抽出することは読者にとって煩雑な作業となっている．そのためそれぞれの関連性や連続性に基いて有用な情報の抽出を行うことでユーザ支援を行う研究が盛んに行われている [1] [2]．また，情報の抽出だけではなく，記述中に登場するキーワードが対応する百科事典上の記事を明示することで，記事の補完を行い理解支援を行う研究についても関心が寄せられている [3]．しかし，多くの研究は記事の報じられ方や記事間の関連性, 分類等に着目しており，ユーザの理解困難な記述について補足する部分を明示する研究は少ない [1] [2] [3].

継続して報道されるニュース記事では，それまでに報道された内容は短い言葉に簡略化し言及されていることや，由来や背景について簡潔に記述されていることが多い．そのような場合，関連性や連続性に基づき記事を推薦する理解支援方法では，推薦された記事が読んでいる記述中のどこに対応するのか対応関係が明示出来ず，読者が理解しにくいと感じた部分に対して情報を補完する効果は薄いと考えられる．

過去に報じられた事象に対し簡潔に記述している例を以下に示す．

“The Park Geun-hye administration is drawing flak for its poor response to the Middle East respiratory syndrome outbreak, despite rising public concerns of the surging number of confirmed or suspected patients. ”(注1)

上記の記事は朴大統領の MERS とよばれる病原体への対応が強く非難されていることに関するものであるが，文章中の ”poor response” という記述が指す事象を理解することは，背景知識の無い読者にとっては容易ではない．また，上記の記述は，以前までに報道された複数の事象をまとめて抽象的に言及しているため，具体的な出来事に結び付きにくい．

本研究では，事象をニュースで報道されている出来事とし，出来事の時間，場所，登場人物が明記されている，一意的に識別できる事象を具体事象とする．これに対して，対応する事象が複数あったり，時間・場所・登場人物があいまいであったり，一意に識別できない事象を抽象事象と呼ぶ．我々は，文章中に記述されている抽象事象について，具体的に誰が何をしたのかを読者が想像することが容易になれば，理解支援になると考え，事象の抽象度の推定手法とそれに基づく事象間の関係を分析する手法を提案する．

本研究の主な貢献を以下にまとめる．

（1） 事象の 5W モデルとその実現手法：文章に記述される事象を Who，Whom，What，When，Where の 5W 要素で表現するモデルを提案する．また，係り受け構造の解析や文章の潜在トピックによるクラスタリングを利用して事象の 5W 要素を抽出する手法の提案を行う．(3. 1,3. 2 節)

（2） 事象の抽象度の推定手法：5W モデルを用いて，事象の 5W 要素を抽出する難易度及びそれぞれの要素の抽象度に基づいて事象の抽象度を推定する．(4. 1 節)

（3） 事象関係の分析手法：5W 要素ごとの階層構造に基づいて事象間の類似度を計算し，抽象度によって類似する事象間の抽象-具体の対応関係を分析する．(4. 2 節)

2. 関連研究

ニュース記事の理解支援

ニュース記事の理解支援に関して，これまでに多く研究されている．

例えば，メディアが持つバイアスに着目し分類を行う研究として，NewsCube [1] では，ユーザーはニュースには多様な側面があるということを認識しなければならず，それを認識することが理解支援につながるとし，Aspect-level を定義，これに従い分類を行うことでメディアのバイアスを最小化し理解支援を行っている．

他のアプローチとして，記事中で述べられているものごとに関

(注1)：KoreaHerald:Blue House blasted for MERS response
http://www.koreaherald.com/view.php?ud=20150603001128

する情報を補完する理解支援方法もある．Rada Mihalcea ら [3] は, Wikify!と呼ばれる，与えられたテキストに対し Wikipedia をリソースとした自動的なアノテーションおよびリンク付与 (wikification) を行うことで読者にキーワードがどのような意味を表すかを提示し，読者の知識補完による理解支援を行っている．

事 象 抽 出

Wikify!のように，読者の持つ興味に着目し情報を補う研究として，NewsStand [2] では, 読者の持つ事象 X がどこで起きたのか，あるいは場所 Y ではどのような事象が起きたのかという疑問を解消するため，ニュース記事中に出てくる単語について Geo taggingn を行い地図上に事象を関連付け表示するインターフェースを開発している．NewsStand は記事単位ではなく，記事中に述べられている事象を対象としている点や，記事中で述べられる単語から地図上の位置を推定している点が我々の研究と類似している．

記事中に述べられている事象の抽出は，ニュース記事の理解支援という目的だけではなく，Event Extraction と呼ばれ，Web 上の記事の内容抽出に関するタスクとして認識されており，事象抽出に関する研究も多く存在する [4].

事象抽出に関する研究としては，Kira ら [5] は，抽出した事象の因果関係を分析することで，与えられた事象がどのような事象を引き起こすかを予測する研究を行っている．彼らは，Kim らの *Property Exemplification of Events theory* [8] を拡張し，事象を引き起こした人物 (Actor) と，何を用いて (Instrument), どのように引き起こしたか (Action)，その対象 (Object) 及び時間 (Time) と場所 (Location) のタプルを事象の表現モデルとして用いている．我々は，事象間の抽象-具体関係を求めるため，事象を特定するための必要な要素として Who, Whom, What 及びそれが発生した時間（When）と場所 (Where) を用いて事象の 5W 表現モデルを提案している。

Web 文書に関する抽象度に着目した研究として，田中ら [6] は，Web 文書の抽象度および具体度について，文書の抽象度，具体度とは単語それぞれが持つ具体度を足しあわせたものの平均として表現している．単語の具体度には，Paivio ら [9] の定義した具象性 (concreteness) および心象性 (imaginability) を用いて人手で分類された，Medical Research Council Psycolinguistic Database(注2) を教師データとして様々な尺度により計算を行っている．田中らの研究では，単語の具体度に着目することで全体としての文章の抽象度を推定しているが，我々が目指すものは，文章そのものではなく，文章中に記述されている事象の抽象度を推定することである．

3. 5W モデル

我々は，*Who*, *Whom*, *What*, *When*, *Where* の 5 つの要素を用いて事象を表現する．これらは，誰が事象を引き起こしたのか (Who)，何が起きたのか（What)，事象の作用対象（Whom)，いつ（When)，どこで起きたのか（Where) によって事象が表現されることを示している．これら 5W 要素により事象 $E = (Who, Whom, What, Where, When)$ と表す．本節ではまずニュース記事から事象を抽出する方法について述べ，次に文章中から 5W 要素が得られない場合の補完方法について述べる．システム処理の概要を図 1 に示す.

3.1 事象の抽出

文章中からの事象抽出方法について述べる．行動や出来事のような事象について記述している文章が与えられた時，その文章の主語 (subject)，目的語（object)，述語 (Verb) をそれぞれ Who，Whom，What に対応付け，時間と場所の表現をそれぞれ When と Where とする．我々は，Stanford Core NLP [10] を利用し，代名詞の照応・共参照解析や固有名詞の抽出，係り受け構造の解析を行う．また，固有名詞として抽出したエンティティに対する多義性の解消のため，AIDA(Accurate Online Disambiguation of Entities) [7] を利用することで固有名詞の曖昧性除去を行う.

3.2 時間・場所要素の補完

我々は，文章中に時間や場所表現が含まれない際の要素抽出の問題解決のため，Latent Dirichlet Allocation(LDA) による文章のトピック分布を利用したクラスタリング行うことで文章と文章が共有する時間や場所表現の推定を行う.

文章から 5W 要素を抽出する際，文章によっては 5W 要素が省略されている場合がある．特に時間や場所については明示的に述べられず，省略されていることが多い．省略されている場合にも，既出であったり，文脈上明らかであるため省略されている場合や，明らかにすることが出来ないために記述していない場合がある．明らかにすることが出来ない場合は，要素の抽象度は最も高いと推定出来るが，文脈上明らかであるため省略されている場合は周辺の文章から推定し補完する必要がある.

そこで我々は同様の文脈の文章は，全て同じ時間や場所を暗黙的に共有していると仮定し，文脈に従い文章をクラスタリングすることで時間や場所が明記されていない文章についても，明記されている文章のものを参照することで推定を行う方法を提案する.

文脈を推定する手法として，文書中に複数のトピックが混在することを仮定する LDA を用いる．LDA を適用する際のパラメータであるトピック数については, 5-fold の交差検証を行い，Perplexity の平均値が最も低いものを採用する.

各文章ごとのトピック分布が得られた後，トピック分布を対象とした *Spectral clustering* によるクラスタリングを行う．Spectral clustering によるクラスタリングでは，クラスタ数 K を決定する必要があるが，前述の方法により自動的に決定されたトピック数をクラスタ数として用いる．尚，トピック分布は多項分布であるため，距離尺度として確率分布の差異を距離関数として扱う．文章 s_1および s_2 の距離 $Dist(s_1, s_2)$ は，正規化したそれぞれの文章のトピック分布 $\boldsymbol{\theta}_{s_1}$，$\boldsymbol{\theta}_{s_2}$ および JensenShannon ダイバージェンス (JS ダイバージェンス) を用いて以下のように計算される.

(注2)：http://www.psych.rl.ac.uk

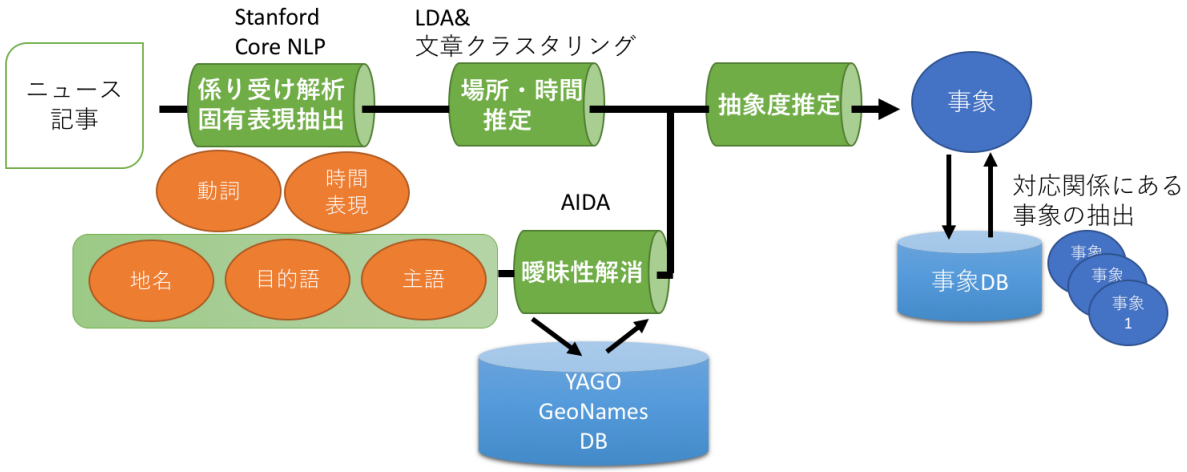

図 1　システムの処理概要

$$Dist(s_1, s_2) = D_{JS}(\boldsymbol{\theta}_{s_1}||\boldsymbol{\theta}_{s_2}) \quad (1)$$

$$D_{JS}(\boldsymbol{\theta}_{s_1}||\boldsymbol{\theta}_{s_2}) = \frac{1}{2}D_{KL}(\boldsymbol{\theta}_{s_1}||M) + \frac{1}{2}D_{KL}(\boldsymbol{\theta}_{s_2}||M)$$

$$M = \frac{1}{2}(\boldsymbol{\theta}_{s_1} + \boldsymbol{\theta}_{s_2})$$

$$D_{KL}(\boldsymbol{\theta}_{s_1}||\boldsymbol{\theta}_{s_2}) = \sum \boldsymbol{\theta}_{s_1}(i) log \frac{\boldsymbol{\theta}_{s_1}(i)}{\boldsymbol{\theta}_{s_2}(i)}$$

もし文章中に時間や場所表現があればそれを利用し，含まれなければクラスタ中の文章の内最も距離の小さい文章中の表現を採用する．ただし，クラスタ中に表現のある文章を含まない場合については，時間表現は，記事の発行日を終端として，記事中にあらわれる最も古い時間表現を始点とする期間を採用し，場所については GeoNames における階層構造を利用し，記事中にあらわれる場所表現の最小共通祖先を採用することとする．図 3 に示すように， 例えば，”New York City”と，”Albany”という場所表現が記事中に含まれている場合，最小共通祖先は”New York State” となる．

4.　事象間関係分析

本節では，事象間の関係分析を行うための方法として事象の抽象度の推定および事象間の関連度の計算方法について述べる．

4.1　抽象度の推定

本来，事象とは特定の時間や場所で起きたものを指す．しかし，文章中の事象への言及には特定の事象だけではなく，時間や場所，主語などに幅をもたせ複数の事象について言及していることがある．そのため，事象の抽象度を，事象を一意に特定する難しさとして定義する．本稿では，以下の二つの側面から抽象度の推定を行う．

（1）　5W 要素それぞれの抽象度

（2）　事象を一意に特定するための 5W 要素の十分さ

5W 要素それぞれの抽象度とは，事象を構成する各要素の言及範囲がどの程度まで及ぶかを推定するものである．例えば，要素の内時間を例にとると，”Congress has not approved major gun-control legislation since the 1990s.”(注3) という文章は，1990-1999 年頃から現在までの複数の事象について言及している．このように，時間として言及している範囲が広くなればなるほど，複数の事象が該当し，抽象度は高くなる．

事象を一意に特定するための要素の十分さとは，要素が複数そろっておりそれらが十分小さい抽象度を持っていれば全ての要素が完全にそろっていない場合でも事象を一意に特定することが可能であることを考慮したものである．例えば，事象の要素の内，特定の日付にのみ起こったことであれば，主語，述語，目的語，場所があれば時間を補完して事象を一意に特定することが出来る．あるいは，特定の場所ではある人物のみが特定の行動をするという事実があれば，主語以外の要素で事象を特定することが可能である．

このように，要素同士は事象を一意に特定するにあたり独立ではなく相補関係にある．そのため，要素がどれだけそろっているか，それらの要素がどの程度の抽象度を持つかを考慮し事象の抽象度を定める．

以下ではまず事象を構成する要素それぞれの抽象度の推定方法について述べ，次にそれらの要素により事象の抽象度を表現する方法について述べる．

4.1.1　主語及び目的語の抽象度推定 (Who, Whom)

事象中の主語及び対象の抽象度を推定するために，オントロジーによる汎化構造の解析を行う．ニュース記事中に登場する主語や目的語の抽象度が異なる例として，下記の例が考えられる．

- Japan foreign minister arranging Seoul visit to settle ’comfort women’ row.
- Japan Foreign Minister Kishida says arranging visit to South Korea.

(注4)

上記は，ほぼ同一の内容に関する記述であるが，”Japan fore-

(注3)：http://www.reuters.com/article/us-usa-obama-guns-idUSKBN0UM0AU20160108

(注4)：http://uk.reuters.com/article/uk-japan-southkorea-idUKKBN0U801M20151225

gin minister”と”Japan Foreign Minister Kishida”のように言及している主語の記述に差がある．主語にだけ注目した場合，人名まで詳しく言及している方がより具体的であると感じられる．

前述した例を考えると，”Kishida”は，”Japan Foreign Minister”というクラスのインスタンスであると考えることが出来，このような汎化構造を利用することでエンティティの抽象度の推定を行う．汎化構造の獲得に際し，我々は Wikipedia や WordNet などを資源として構築された知識ベースである YAGO を利用する．YAGO では，2012 年の時点で 1000 万に及ぶエンティティ及び，エンティティ間の関係を表す fact1 億 2000 万が蓄えられている．エンティティ間の関係性を表す fact は，(エンティティ, relation, エンティティ) のように三つ組で表現されており，例えば，(Shinzo Abe, type, Goverment minister of Japan) といった固有名詞と’type’ 関係にある汎化クラスが得られる．

また，クラスには，階層構造があり，上位のクラスが存在する．エンティティであれば最も上位のクラスは ”Entity” となる．文章中で主語や目的語となる単語はほぼ全て固有名詞であると考えられるため，最上位クラスである ”Entity” の抽象度を 1 とし，位置する階層が深くなるにつれ減衰する関数によって抽象度の推定を行う．

5W 要素の抽象度について要素 c の深さ $depth(c)$ に応じて非線形に減衰する抽象度関数 $abst_{Depth}$ を以下のように定義する．

$$abst_{Depth}(c, \lambda) = \exp\left(-\frac{depth(c) - 1}{\lambda}\right) \tag{2}$$

減衰の割合を調整する定数項 λ を導入し，階層によって非線形に減衰する関数を定義した．これは，YAGO 上では階層の差が 1 である，”Person”と”Government ministers”間に比べ，”Government ministers”と”Shinzo Abe”間の抽象度の差を同様であるとは考えにくく，ニュース記事中の言及の抽象度とは，線形に推移するのではなく非線形であると考えられるためである．

主語 sbj，目的語 obj の抽象度 $abst_{sbj}$ および $abst_{obj}$ は (2) 式により以下のように定義される．

$$abst_{sbj}(sbj) = abst_{Depth}(sbj, \lambda_{sbj}) \tag{3}$$
$$abst_{obj}(obj) = abst_{Depth}(obj, \lambda_{obj}) \tag{4}$$

定数項 λ は，5W 要素のそれぞれで減衰率が異なると考えられるため， λ_{sbj} および λ_{obj} として別々に定義している．

4.1.2 動詞の抽象度推定 (What)

What を表す動詞については，動詞の属する意味フレームの階層構造及び語が属する意味フレームの個数を利用し抽象度を計算する．意味フレームとは，FrameNet(注5) と呼ばれる語彙データベースに含まれる，動詞の持つ意味的な役割を，想起されるエンティティと共に表現したものである．例えば，cooking という単語の持つ概念は，料理する人であったり，料理そのものや，加熱するための器具を含む．FrameNet では，器具や料理人のような喚起されるエンティティを FrameEntity と呼び，概念を Frame，概念に相当する単語を LexicalUnit と呼んでいる．通常，動詞が LexicalUnit として属する Frame は複数ある．例えば最も属するフレームが多い動詞は”rise”であり，”Motion direction” や ”Rising to a challenge” などが Frame としてあげられる．また，Frame 同士には継承と呼ばれる Frame 間の関係が定義されており，これを Frame の階層構造として利用する．Frame の階層構造の例を図 2 に示す．

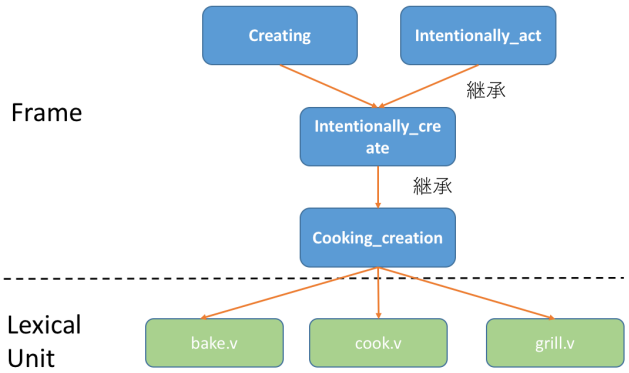

図 2　Frame 階層構造の例

我々は，対象の動詞 v を LexicalUnit として含むフレーム集合を F_v として， フレーム集合に属するフレーム $f_i \in F_v$ の個数を $|F_v|$ ，フレームの深さを $depth(f_i)$ と定義する．また，前述した $abst_{Depth}$ 関数を用いてフレーム集合 F_v に対する階層に関する抽象度 $Depth_{F_v}$ を以下のように定義する．

$$Depth_{F_v} = \frac{1}{|F_v|} \sum_{f_i \in F_v} abst_{Depth}(f_i, \lambda_{verb})$$

さらに，所属するフレームが 1 個のとき抽象度が 0 になり，属するフレームが最多である”rise”の属するフレーム数 9 が 1 となる所属フレーム数による抽象度を定義し，階層に関する抽象度との重み付き和として動詞 v の抽象度 $abst_{verb}(v)$ を以下のように定める．

$$abst_{verb}(v) = \alpha(1 - \exp(-\frac{|F_v| - 1}{2})) + (1 - \alpha)Depth_{F_v} \tag{5}$$

α は定数項であり．今回は $\alpha = 0.5$ として階層に関する抽象度及びフレーム個数による抽象度の平均値を動詞の抽象度として用いることとする．

4.1.3 場所の抽象度推定 (Where)

Location の抽象度推定のために，我々は地理情報データベースである GeoNames(注6) を利用する．GeoNames には，1,000

(注5)：https://framenet.icsi.berkeley.edu/fndrupal/

(注6)：http://www.geonames.org

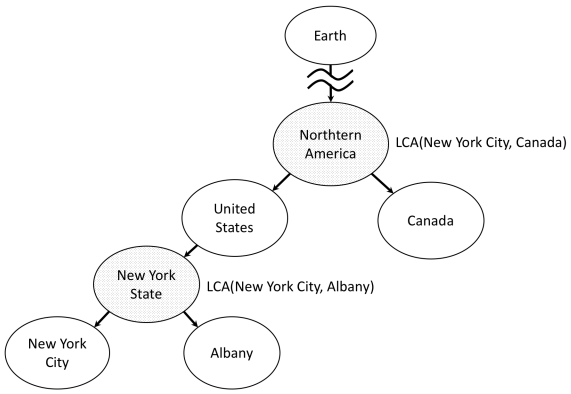

図 3　GeoNames から得られた地名の階層構造の例

万を超える地名が含まれており，それぞれの地名に国や町などのクラスが割り当てられている．また，”Albany City”が存在するのは”New York State”であるといった親子関係も保持しており，これを利用することで地名の階層的位置づけを得る事ができる．最も上位に位置する ”Earth” を基点とし，”Continent”, ”Countroy”と階層的に深くなるほどより具体的になると考える．GeoNames によって得られる地名の階層構造の例を図 3 に示す．

また，事象中の時間，場所は，文章中から抽出したものと，4.1 節にて行った推定によって得られる場合がある．推定によって得られた時間や場所については抽象度がより高くなると考えられる．そのため，事象が抽出された文章 s_1 と，推定時に参照する文章 s_2 との距離 $Dist(s_1, s_2)$ を用いて，距離が遠い文章を参照する程より抽象度を高くする．

以上から場所 loc の抽象度 $abst_{loc}$ を以下のように定義する．

$$abst_{loc}(s_1, s_2, l) = Dist(s_1, s_2)(abst_{Depth}(l, \lambda_{loc}) - 1) - abst_{Depth}(l, \lambda_{loc}) \tag{6}$$

4.1.4　時間の抽象度推定 (Time)

When を表す時間の抽象度については，期間の長さにより計算する．例えば，”2016 年 3 月”という表現よりも，”2016 年 3 月 1 日”という言及のほうがより具体的である．抽出する時間の最小単位を日とし，時間の長さを t としたとき，事象が抽出された文章 s_1 と，推定時に参照する文章 s_2 との距離 $Dist(s_1, s_2)$ を用いて時間の抽象度 $abst_{time}$ を以下のように定める．

$$abst_{time}(s_1, s_2, t, \lambda_{time}) = \exp\left(-\frac{t-1}{\lambda_{time}}\right)(1 + Dist(s_1, s_2)) - 1 \tag{7}$$

ここで λ_{time} とは，階層構造の抽象化時に利用したものとは対照的に，増加の割合を調整するための定数である．

4.1.5　事象の抽象度推定

最後に事象の抽象度推定を行う．事象の抽象度推定には，事象を一意に特定するための 5W 要素の十分さの側面から，これまで求めた 5W 要素それぞれの抽象度を用いる．

前述したように，5W 要素はそれぞれ独立に存在するわけではなく，相互に密接に関係しており，ある要素が欠けている場合においても他の要素が十分に低い抽象度を持っている場合には推定が可能になる．また，要素それぞれについても一意に特定可能な程度は異なっていると考えられる．例えば，事象に人物が含まれていれば，普段の人物の行動から場所や行動はある程度推定可能になる．このように，要素ごとに事象の一意な特定可能性は異なっており，これを考慮する必要がある．

以上から，上述した特定の可能性を 5W それぞれの重みベクトル $\boldsymbol{w} = (w_{sbj}, w_{obj}, w_{verb}, w_{loc}, w_{time})$ として表現し，事象の抽象度を線形和として表現する．尚，事象の要素集合を $E = \{sbj, obj, verb, loc, time\}$ として表し，事象 $\boldsymbol{e_i}$ は $\boldsymbol{e_i} = (e_{i,sbj}, e_{i,obj}, e_{i,verb}, e_{i,loc}, e_{i,time})$ のように構成される．

$$abst_{Event}(e_i) = \frac{1}{|E|}\sum_{c \in E} w_c abst_c(e_{i,c}) \tag{8}$$

4.2　事象間の対応関係

事象間の抽象・具体関係分析には，話題や文脈が共有されている，対応関係にある事象の組の抽出と，事象毎に推定した抽象度を用いてどちらがより抽象的かの判別の二つの段階がある．抽象度は個々の事象の要素ごとに独立して計算を行うが，得られた抽象度は事象が属する話題毎に取りうる値の範囲が大きく異ると考えられる．そのため，同一の話題に関する事象の組を対応しているとし，対応関係にある事象間でのみ事象の抽象度を指標として用いることとする．事象間に存在する関係性を表現する特徴の抽出を行い，SVM を用いて事象の対応関係の有無を判別する．

これまで要素ごとの階層構造について言及したが，この階層構造を利用すると，$c \in E$ となる事象 e_i の要素 $e_{1,c}$ が事象 e_j の要素 $e_{2,c}$ の子孫である場合，意味的に包含するため，$e_{1,c} \subseteqq e_{2,c}$ のように部分集合であると表す事ができる．この意味的に包含する関係を利用し，包含関係であれば 1，さもなければ 0 といったように真偽値のベクトルを作成する．

要素間の祖先-子孫関係を利用し，$e_{i,c}$および $e_{j,c}$ の要素間に祖先-子孫関係がある場合に 1 となる関数を以下のように定義する．

$$\delta_{e_{i,c}, e_{j,c}} = \begin{cases} 1, & \text{if } e_{i,c} \subseteqq e_{j,c} \quad or \quad e_{j,c} \subseteqq e_{i,c} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

尚，時間表現については期間として $t2$ が $t1$ の期間を完全に含む関係であれば $t_1 \subseteqq t_2$ としている．また，主語，目的語については，固有エンティティ単体の属性に関する階層構造をあらわしてはいるが，例えばオバマ大統領がアメリカ合衆国に属している，といった所属関係は得られないため，YAGO 上の fact の述語によるリンク関係を利用し，s_1 から s_2 にリンク関係がある場合，$s_1 \subseteqq s2$ もしくは $s_2 \subseteqq s_1$ であるとみなし，$\delta_{s_1, s_2} = 1$ とする．抽出された事象の要素間に関する対応関係の例を図 4

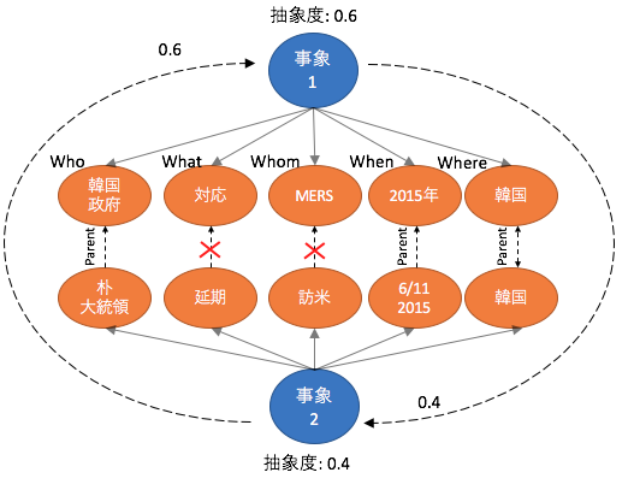

図 4　事象間対応関係の例

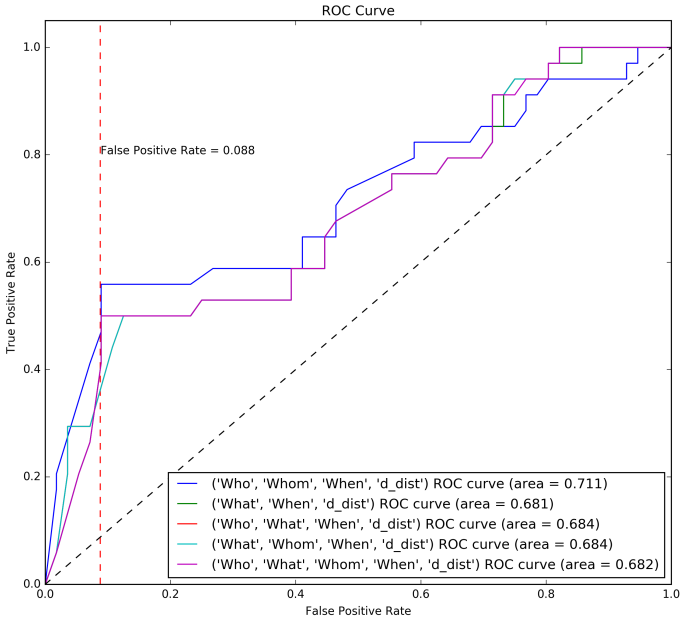

図 5　上位 5 つまでの組み合わせによる ROC 曲線

に示す.

事象間の要素の祖先・子孫関係に加え，事象 e_1, e_2 が抽出された文書 d_1, d_2, および文章 s_1, s_2 のトピック分布間の距離 式 (1) を利用しそれぞれ $Dist(d_1, d_2), Dist(s_1, s_2)$ のように表現し，特徴として用いる．文章，文書のトピック分布導出には，4.1 節と同様に 5-fold の交差検証にて最も低くなる perplexity を算出しトピック数を決定する.

上述した，要素間の意味的な包含関係および文章，文書トピック分布の距離の組み合わせから，事象間の対応関係を分類するにあたり有用な特徴を選定する.

5. 評 価 実 験

5.1 対応関係の評価

評価実験のため，複数のニューストピックからそれぞれ一定期間内でランダムに抽出した記事中の事象間についてデータセットの作成を行う．今回実験で用いたニューストピックは以下のとおりである.

- Volkswagen の CO2 排出量に関する報道
- アメリカ大統領選挙に関する報道
- 安倍首相に関する報道

また，上記に関する記事は，メディアごとの記事の書き方の影響を避けるため，全て The New York Times(注7) から取得している.

これらのニューストピックの記事から事象を抽出し，10 個のクエリ事象とそれぞれのクエリの事象の対応事象候補を 30 個を選び，クエリ事象と候補事象のペア，計 300 ペア間の対応関係の有無を判定する．具体的に，以下のステップで事象ペアを作成する.

（1） ニューストピックごとに三つの記事を選ぶ.

（2） 全記事中から事象の抽出を行う.

（3） 得られた全事象から無作為に 10 個のクエリとなる事象を選ぶ.

（4） クエリ事象と主語に対応のある事象 30 個を全事象から無作為に選び，事象ペアとする

ステップ (4) にて主語に対応のある事象のみを選択したのは，完全に無作為に選んだ場合ほとんどが対応関係無しとなるのを防ぐためである．それぞれのニューストピックから得られた事象および事象の言及元記事タイトルの例を 表 1 に示す.

事象ペアの対応関係有無の判定基準として，クエリ事象と候補事象の一方が他方の具体的な行動となる場合，あるいは非常に似ている事象の場合に対応関係があるとして判定を行った．分類の結果，対応関係のある事象の組はデータセット全 300 組中 74 組となった．分類したデータを元に，線形カーネルの SVM を分類器として特徴を組み合わせ 5-fold の交差検証を行い，ROC AUC の平均および標準偏差を計算した結果上位 10 件を表 2 に示す.

表中，特徴の組に各特徴の組み合わせを示しており，d-dist は文書間距離，s-dist は文書間距離である．また，AUC が上位 5 位までの特徴の組み合わせを入力とし，データセットの 7 割を訓練データ，3 割をテストデータとしたときのそれぞれの ROC 曲線を図 5 に示す.

表 2 から，最も AUC の高い組み合わせは，(Who, Whom, When, 文書間距離) の四つの特徴の組み合わせである．また，上位の ROC 曲線を示した図 5 は，False Positive Rate が低い際に他の組み合わせより True Positive Rate が顕著に高くあらわれることが示している.

上位 5 位までの組み合わせ中には共通して When および文書間距離が含まれており，分類にはこの二つの特徴が特に有効であることがわかる．尚，データセット作成の際に，Who が対応関係にある事象を対象としているため，Who が対応関係にある事象群では What, Whom, When，文書間距離が分類に有効であり，Where や文章間距離は分類に比較的寄与していないということが考えられる.

(注7)：http://www.nytimes.com

表 1　ニューストピック毎の事象の例

ニューストピック	主語	動詞	目的語	場所	期間	言及元記事タイトル
Volkswagen スキャンダル	Volkswagen	halt	sales	United States	2015-09-20 - 2015-09-20	Volkswagen to Stop Sales of Diesel Cars Involved in Recall
大統領選挙	Donald Trump	created	pool	Iowa	2015-06-01 - 2015-09-30	2nd-Place Finish Pierces Donald Trump's Mystique, but Another Chance Comes Quickly
安倍首相に関する報道	Shinzo Abe	told	Parliament	Japan	2015-01-19 - 2015-01-25	Departing From Japan's Pacifism, Shinzo Abe Vows Revenge for Killings

表 2　各特徴の組み合わせを入力とした SVM による分類結果

特徴の組	ROC AUC	標準偏差
Who,Whom,When,d-dist	0.680430	0.072135
What,When,d-dist	0.663564	0.069450
Who,What,When,d-dist	0.660500	0.062955
What,Whom,When,d-dist	0.660500	0.062955
Who,What,Whom,When,d-dist	0.653965	0.076205
When,s-dist	0.652335	0.107128
Whom,When,s-dist	0.634689	0.122228
Whom,s-dist	0.626186	0.088462
Who,s-dist	0.626186	0.088462
Who,When,s-dist	0.625353	0.131786

表 3　分類に失敗した事象例

	主語	目的語	動詞	場所	期間
1	Donald Trump	did	None	Iowa	2016-02-01 - 2016-02-01
2	Donald Trump	focusing	None	Iowa	2016-02-01 - 2016-02-01

分類に失敗した例として，以下のようなものがあげられる．

（1）“And when he indulged in the pandering to Iowa institutions that is typical of political supplicants here, he did so in his exaggerated, almost comic style – as if he were playing the role of presidential candidate .”

（2）“But other said they think Trump should be focusing on the next contests.”

これはそれぞれ，“Trump Calls for Iowa Election Do-over”[注8]，および “Ted Cruz Wins Republican Caucuses in Iowa”[注9]．の記事中から得られた事象である．これらから抽出された事象を表 3 に示す．

表を見ると，主語，場所，時間が対応関係にあり，対応関係ありとして分類結果が得られた．しかし，元の記述では ”They think” として文章が始まっており，これは筆者の主観的な意見が述べられていることがわかる．そのため，実際に起きた (1) の事象とは対応関係無しとラベル付けされ，分類に失敗している．

(注8)：http://www.nytimes.com/aponline/2016/02/03/us/politics/ap-us-gop-2016-trump-iowa.html

(注9)：http://www.nytimes.com/2016/02/02/us/ted-cruz-wins-republican-caucus.html

表 4　抽象度によるランキングの評価結果

セット番号	nDCG@ 5	nDCG@10	nDCG@15
0	0.583513	0.624021	0.672747
1	0.324303	0.410066	0.534586
2	0.559329	0.558242	0.600818
3	0.382865	0.535759	0.649355
4	0.342492	0.331362	0.422276
5	0.437096	0.441723	0.504354
6	0.476428	0.598271	0.622410
7	0.703007	0.696193	0.704059
8	0.640435	0.636034	0.675917
9	0.532158	0.544703	0.663410

表 5　ニューストピックごとのランキング評価結果

ニューストピック	nDCG@ 5	nDCG@10	nDCG@15
Volkswagen	0.327385	0.401008	0.461643
Election	0.677131	0.603552	0.586908
Shinzo Abe	0.411833	0.429399	0.464184

以上の結果から，対応関係の分類候補事象からは主観的な記述を除く前処理が必要であることがわかる．

5.2　抽象度による順序関係の評価

次に，対応関係にある事象間の抽象度による順序関係の評価を行う．

節 5.1 にて作成したデータセット中，クエリ事象を含めた候補事象との 31 事象を 1 セットとし，全 10 セットを抽象度による順序関係評価用データセットとした．順序関係評価用データセット中それぞれについて，事象を抽出した元の記述に人手による抽象度の点数付けを行った．点数は 1 点から 5 点までの 5 段階評価とし，事象が抽出された元の記述が対応する事象が多ければ抽象度が高いと判断し，点数を高く付与した．

節 4.1 にて行った事象の抽象度によるランキングの nDCG による評価結果を表 4 に示す．

4 を見ると，nDCG@5 の結果は各セットごとに大きく異なっていることがうかがえる．この要因を突き止めるため，我々は次にニューストピックごとの nDCG の評価を行った．結果を表 5 に示す．表を見ると，抽象度によるランキングの評価結果は，セットごとではなくニューストピックごとに差異があることがわかり，アメリカ合衆国の選挙に関する事象が最も高く，Volkswagen に関する事象が最も悪いという結果を示している．

表 6　抽象度の点数が最も異なる例

抽象度	点数	主語	目的語	動詞	場所	時間
0.32	4	Volkswagen	surpassed	Toyota	New York	2015-10-26

今回，最も低い値となった Volkswagen に関する事象の内，抽象度と人手による点数が最も異なる例を以下に示す．表中，抽象度はシステムにより計算された値であり，点数が人手により付与された抽象度の点数である．

表 6 に示した事象は Volkswagen に関する事象の内最も抽象度の低いものとしてランキングしているが，人手による点数では抽象度が高いことを示している．事象の抽出元記述では，

”Only three months earlier, Volkswagen had surpassed Toyota for the first time and, however briefly, realized a long-held dream to become the world ’s biggest automaker.”

のように，3ヶ月の間という注釈があるが，抽出された事象では期間が 2015 年 10 月 26 日の 1 日間となっており，期間の推定が正しく行えていないことがわかる．今回扱った Volkswagen の記事や安倍首の記事では，過去への言及部分と，報道された日時に起きた事象への言及が混じったものが多く，それらの個々の事象の期間の推定が正しく行えていない事例が散見された．一方，比較的よい結果となったアメリカ合衆国の大統領選挙に関する事象では，記事中から抽出される事象のほとんどが同一の期間であることが多く，選挙報道の性質として過去への言及があまりされないという傾向があることが考えられる．

6. おわりに

本稿では，ニュース記事中の文章から事象を抽出し表現する 5W モデル，5W をベースとした事象間の抽象-具体の対応関係を分析する手法を提案した．また，事象の対応関係分類および抽象度によるランキングに関する評価実験を行った．

対応関係分類の評価実験では，対応関係の分類に有効な特徴量として Who, Whom, When, 文章間距離が得られた．分類ができていなかった事例では，対応関係の候補事象について筆者の主観的記述を対象外とする前処理の必要性が示された．

また，抽象度によるランキングの実験結果にて，抽象度の推定に誤りがあった事例として事象の時間推定に関する問題が示された．今後，5W 要素の補完精度の評価および大規模な実験による評価が課題としてあげられる．

謝　辞

本研究の一部は，科研費（課題番号 25700033) による．

文　献

[1] Park, Souneil, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. ”NewsCube: delivering multiple aspects of news to mitigate media bias.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 443-452. ACM, 2009.

[2] Teitler, Benjamin E., Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. ”NewsStand: A new view on news.” In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems, pp. 18. ACM, 2008.

[3] Mihalcea, Rada, and Andras Csomai. ”Wikify!: linking documents to encyclopedic knowledge.” In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 233-242. ACM, 2007.

[4] Liao, Shasha, and Ralph Grishman. ”Using document level cross-event inference to improve event extraction.” In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 789-797. Association for Computational Linguistics, 2010.

[5] Radinsky, Kira, Sagie Davidovich, and Shaul Markovitch. ”Learning causality for news events prediction.” In Proceedings of the 21st international conference on World Wide Web, pp. 909-918. ACM, 2012.

[6] Tanaka, Shinya, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. ”Estimating content concreteness for finding comprehensible documents.” In Proceedings of the sixth ACM international conference on Web search and data mining, pp. 475-484. ACM, 2013.

[7] Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Frstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. ”Robust disambiguation of named entities in text.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 782-792. Association for Computational Linguistics, 2011.

[8] Kim, Jaegwon. Supervenience and mind: Selected philosophical essays. Cambridge University Press, 1993.

[9] Paivio, Allan, John C. Yuille, and Stephen A. Madigan. Concreteness, imagery, and meaningfulness: Values for 925 nouns. American Psychological Association, 1968.

[10] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. ”The Stanford CoreNLP natural language processing toolkit.” In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60. 2014.